Environment, Development and Sustainability (2023) 25:11209–11239
https://doi.org/10.1007/s10668-022-02524-y

Understanding compliance with voluntary sustainability standards: a machine learning approach

Anja Garbely[1] · Elias Steiner[1]

Received: 28 April 2021 / Accepted: 9 June 2022 / Published online: 4 July 2022

Abstract
Voluntary sustainability standards are quickly gaining ground. Whether and how they work in the field, however, remains largely unclear. This is troubling for standards organizations since it hinders the improvement of their standards to achieve a higher impact. One reason why it is difficult to understand the mechanics of VSS is heterogeneity in compliance. We apply machine learning techniques to analyze compliance with one particular VSS: Rainforest Alliance-for which we have detailed audit data for all certified coffee and cocoa producers. In a first step, we deploy a *k-modes* algorithm to identify four clusters of producers with similar non-compliance patterns. In a second step, we match a large array of data to the producers to identify drivers of non-compliance. Our findings help VSS to implement targeted training or risk assessment using prediction. Further, they are a starting point for future causal analyses.

Keywords Voluntary sustainability standards · Machine learning · Compliance

JEL Classification O13 · Q10 · C01

1 Introduction

In the past decades, consumers in rich countries have increasingly taken interest in the social and environmental aspects of the products they buy. This manifests in the rising demand for agricultural products that carry labels such as *Fairtrade*, *Organic*, or *Rainforest Alliance*-so-called voluntary sustainability standards (VSS). Hainmueller et al. (2015), for example, report an average annual growth rate in the demand for Fair Trade certified products in the U.S. of 40% between 1999 and 2008. Such labels are designed with the aim to improve the lives of the producers in the global south or to reduce the impact of the production process on the environment. To achieve this, the standards organizations (SO) behind the labels establish a catalog of criteria according to which goods have to be

✉ Anja Garbely
anja.garbely@unilu.ch

Elias Steiner
eliassteiner@outlook.com

[1] Department of Economics, University of Lucerne, Frohburgstrasse 3, 6002 Lucerne, Switzerland

produced and/or traded in order to obtain their label. The certified producers are usually monitored by independent auditors. If the audit is passed, their product may carry a label signaling to consumers that they are purchasing a good that was produced in line with the standard.

Through the improvement of agricultural practices, VSS have a large potential to contribute to sustainable development in many areas. However, the scientific literature still remains largely inconclusive on the effects of certification (see, e.g., Oya et al., 2018). Most studies try to assess the overall effect of certification by comparing certified to non-certified producers in a particular area while somehow controlling for selection bias (see, e.g., Ruben and Fort, 2012; Waarts et al., 2014; Ingram et al., 2017). As is highlighted in a detailed literature review by Oya et al. (2018), both, the findings and the quality of most studies are mixed or limited. One reason why it is difficult to identify the effect of certification could be compliance-or more specifically: the variance in the degree of compliance with a standard's criteria among producers. Full compliance is usually not necessary to obtain a label. Thus, the effect of certification might depend on a producer's degree of compliance. To give an example, if Ethiopian coffee producers frequently do not comply with environmental criteria, a study looking at whether the certification has a positive effect on environmental outcomes for Ethiopian coffee growers is thus likely to find no effect.

While not attempting to find direct causal evidence for such heterogeneous treatment effects, we are contributing to the literature by investigating compliance with VSS in detail. In particular, we study audit data provided by Rainforest Alliance (RA), one of the biggest global players in the market of VSS. The data are at the certificate level where certificate refers to coffee and cocoa producers that can either be certified as individual farms or groups, such as cooperatives. Among the criteria to obtain RA certification, many are likely to be interrelated. Hence, the hypothesis we aim to verify is that it is possible to identify patterns of non-compliance within the complex structure of the audit data. If there are clusters of producers with similar non-compliance patterns, they are likely to also be comparable along other dimensions. Thus, we state a second hypothesis that it is possible to identify potential drivers of cluster affiliation. Identifying clusters of producers could help to develop tailored training courses. Additionally, finding clusters that struggle to comply in similar areas of the standard is the starting point for the analysis of what is driving non-compliance. A descriptive analysis sheds light on the type of producers that do not comply with different areas of the RA standard. To the best of our knowledge, we are the first to present detailed descriptive evidence on compliance with a VSS on a global level. Besides laying the foundation for causal inference studies or for classification, such information is in itself interesting since it might tell us something about who is having a hard time keeping up with the standard and who does not. This information can help to improve the standard through systematic assignment of training units or ex-ante risk assessment. Finally, as an additional exercise, we check whether it is possible to predict cluster affiliation based on producer characteristics. This sheds light on which variables have the most power in sorting producers into different pre-defined clusters.

To test these hypotheses, we take advantage of recent advances in machine learning. While machine learning techniques have been used extensively in recent years in other fields, it is a rather novel approach for research regarding VSS. One of the main advantages of the methods is that they manage to uncover patterns and structure in complex data unsupervised. We regard this as highly advantageous in our setting where relying on previous theory or empirical results is hardly possible.

Finding structure in the extensive audit data is the first step to understanding compliance. Hence, we first propose an adapted *k-modes* clustering algorithm to uncover frequent compliance patterns and to group producers according to the criteria they do not comply with. We then use a large array of additional indicators-matched to producers using location data-to find common characteristics of certificates within each of these clusters. As a last step, we show the results of a classification exercise where we try to predict cluster affiliation based on these characteristics.

We can identify four clusters: compliers, non-compliers with environmental criteria, non-compliers with management criteria, and non-compliers with social criteria. We find among other insights that producers in the first cluster-the compliers-are more likely to be large individual farms that have been certified for longer and are found predominantly in more developed regions of Central and South America. Producers in the second cluster-non-compliance with environmental criteria-tend to be located in the least developed regions. This cluster contains predominantly small coffee producers that are certified in groups, especially in Ethiopia. Weak management is often found in cocoa production in West Africa-most Ghanaian producers are assigned to this cluster. Finally, group certificates for coffee production in Central America, mainly in El Salvador, are over-represented in the last cluster where issues with social criteria are prevalent.

The paper proceeds as follows. The next section summarizes the existing literature, followed by Sect. 3 outlining the data. The compliance patterns found through clustering are presented in Sect. 4. Section 5 characterizes the clusters, and Sect. 6 shows the results of the classification exercise. After a short summary of our results, we discuss the regarding limitations and existing literature in Sect. 7. This is followed by the implications of the results, and in Sect. 9, we conclude.

2 Literature

As indicated in the previous section, the literature on compliance with VSS is extremely sparse. We know of merely three papers addressing the issue. Kirumba & Pinard (2010) look at UTZ certification of coffee farmers in Kenya. They compare compliant farms to those that never achieved certification although they tried. The results point out that economic drivers determine compliance. However, the study does not provide any details on what the issues with compliance were, they merely suggest a positive selection bias arising from the certification process. The other two studies look at a previous version of the standard by Rainforest Alliance for Brazilian coffee producers only. Pinto et al. (2014) compare the compliance of group to that of individual farm certificates. They find that the two types of producers obtain similar compliance levels. In a more recent paper, Maguire-Rajpaul et al. (2020) in large parts confirm these findings using more data. They find that Brazilian group certificates have slightly higher non-compliance with a selection of management criteria and somewhat lower social performance than individual farm certificates-although the differences are statistically insignificant. Our study extends the analysis to the entire world and more crops. We use more recent data and also provide more structure to the analysis by first clustering the certificates according to their non-compliance patterns, while the previous studies simply compute compliance scores. Our paper relates further to the empirical literature on the impact of certification in general (e.g., Van Rijsbergen et al., 2016; Cramer et al., 2017; Dragusanu and Nunn, 2018; Glasbergen, 2018; Krumbiegel et al., 2018; Borsky and Spata, 2018; Sellare et al., 2020; Dietz et al., 2020). We also contribute to the growing literature on machine learning techniques used in social sciences (e.g., Mullainathan and Spiess, 2017; Chalfin et al., 2016; Kleinberg et al., 2018) and in particular, we add to the methodological literature on clustering algorithms (e.g., Huang, 1997; Khan and Ahmad, 2013; Cao et al., 2013).

3 Data

At the heart of the analysis are audit data at producer level that are collected by third-party auditors and processed by RA. The dataset reports the compliance status for all criteria of the 2017 Sustainable Agriculture Standard, which is one of the most widely adopted VSS. It was developed by RA and the Sustainable Agriculture Network (SAN) and is made up of 119 criteria grouped into four principles: effective planning and management system, biodiversity conservation, natural resource conservation, and improved livelihoods and human well-being. The standard consists of critical and continuous improvement criteria. Compliance with the critical criteria is mandatory, both, for the initial certification and the retaining of certification. Critical criteria are the foundation of the standard and include labor, social, and environmental issues of highest priority and risk. After the initial certification, the standard defines a sequential progress including three levels: C, B and A that consist of the continuous improvement criteria. Every year an increasing share of the continuous improvement criteria must be complied with to retain certification. As the new standard has only been implemented in 2017 and every producer was then set to "year zero", we will not look at the criteria of levels A and B, as no producer had to comply with these at the time of the audits in our data.

The dataset contains the results of 919 independent audits between 2017 and 2019 for 561 certified producers.[1] To reduce the data to one audit per producer, we choose the one in year 2018 where we have most observations. If not available, we use the year 2017 or 2019. Of the producers included in the data, 58.6% are group certificates, 39.2% are individual farms, and a meager 2.1% are multi-site certificates. We count the latter as individual farms to avoid having a very small group. Further, 86.8% produce coffee (Arabica), the most widely certified crop worldwide, and 13.2% produce cocoa, another important crop in the certification business.

The audit result for each criterion can either be compliance (1), non-compliance (0), or non-applicable (NA). Non-applicable criteria refer to practices irrelevant to a producer, e.g., "safe application of pesticide by aircraft" if a producer does not use aircraft. To give an impression of the distribution of audit responses: over all criteria and observations 72.2% are compliant, 7.4% are non-compliant, and 20.4% are non-applicable. Further, the share of NA ranges from 0 to 92.2% depending on the criteria. Of the 561 observations, a total of 91.4% pass the audit and are (re)certified. If a producer does fail the audit, it is predominantly due to non-compliance with one or more critical criteria. Compliance with level C criteria ranges from 50% of the criteria (which is sufficient for certification in the first year) to 100%.

Besides audit data, RA provides data on the individual certificates, which we were able to match to the producer-level audit data. We use variables on the number of workers, the number of producers in group certificates, the operation size in hectares, output per hectares, and the year a producer was first certified.[2] In addition to these variables, the data

[1] After we drop one outlier where we suspect grave measurement error.

[2] Even though the standard was renewed in 2017, many producers were already certified before, according to the previous standard.

contain coordinates of each producer, which we use to match a large array of GIS-coded variables to our data. We use data on the harvested area and quantity of coffee and cocoa in the surroundings of producers, certification density of the two crops by other VSS, population density, distance to the nearest city, child mortality, nighttime light density, vegetation, protected areas and terrain ruggedness. We include a buffer around the producers with a radius of 15km for several reasons: to reduce noise, to control for the fact that the resolution of the GIS data differs depending on the variable, and because some of the group certificates report the location of their city offices. Within this buffer, we aggregate the underlying grid-cells using the mean.[3] Additionally, we look at a number of economic and political variables at the country level. We will present summary statistics in Sect. 5 below. Appendix A provides details of all variables used in the analysis.

4 Compliance patterns

4.1 Method

One of the most popular unsupervised machine learning techniques is the *k-means* algorithm that groups data into similar clusters. For our cluster analysis, we use an adapted version of this algorithm that takes into account the peculiarities of our dataset. Hence, we program our own clustering algorithm using the *C*++ programming language and incorporating findings from the literature. We follow Huang (1997) who introduces the *k-modes* algorithm, as the k-means algorithm is designed for numerical data. As described above, our audit data are binary, in which case both algorithms work in theory. However, we find that k-modes perform better in clustering our particular data.

Additionally, our dataset has a lot of "missing" values since not every criterion applies to every observation. Most observations have some non-applicable criteria; thus, we cannot drop them and we need an algorithm that runs well even with "missing" data. With k-modes, it would potentially be possible to treat non-applicable as an own category. However, we do not want the clustering to be driven by the non-applicable criteria and, thus, rule out this approach. In the following, we describe a version of k-modes that handles this data structure.

To determine the similarity of observations, we need to define a distance measure. For our binary data, we use the *Hamming Distance* that counts the number of dissimilarities between two objects. More specifically, each criterion takes on one of three values in $\{0, 1, \mathrm{NA}\}$, where 0 and 1 stand for non-compliance and compliance, respectively, and NA stands for non-applicable. Thus, the distance between two vectors X and Y (all criteria of two observations) is given by

$$d(X, Y) = \frac{1}{m^{0,1}} \sum_{j=1}^{m} \delta(x_j, y_j)\omega_j, \tag{1}$$

where ω_j is the weight of criterion j described below, and $m^{0,1}$ is the number of criteria that are not NA in either X or Y. As for the distance measure, it is computed as

[3] Note that all cells that have their centroid lying within the 15 km radius are included in the aggregation.

Fig. 1 Clusters *Note*: The first panel of this Figure shows for each cluster (vertical axis) the modes of each criterion (horizontal axis). Panels 2 to 5 show the compliance status for each observation (vertical axis) within a given cluster. Each of the panels 2 to 5 corresponds to one line in the first panel. For example: In cluster NC-M (panel 4), a majority of observations do not comply (red) with criterion 1.7, and hence, the corresponding cell (line 3, column 1) in panel 1 shows a mode ”non-compliance“ for this criterion. Remember that cluster C stands for compliance, NC-E stands for non-compliance with environmental criteria, NC-M stands for noncompliance with management criteria, and NC-S stands for non-compliance with social criteria ▶

$$\delta(x_j, y_j) = \begin{cases} 1: & x_j \neq y_j \quad \cap \quad x_j, y_j \neq NA \\ 0: & x_j = y_j \quad \cup \quad x_j = \mathrm{NA} \quad \cup \quad y_j = \mathrm{NA} \end{cases}. \tag{2}$$

Hence, $d(X, Y)$ is the number of criteria in which $x_j = 0$ and $y_j = 1$ or vice versa, divided by the number of criteria in which $x_j, y_j \neq$ NA. Or, in other words, we do not count NA as different from either 1 or 0. The vector ω represents weights that are proposed by Huang (1997). These weights ensure that homogeneous criteria get more weight for the clustering such that outliers in certain categories are well-detected. The algorithm proceeds in the following steps.

K-Modes Algorithm

```
[1.] Input: $k$ = number of clusters; $D = n$ x $m$ array (data); $M = k$ x $m$ array of initial modes
[2.] Compute the distance of each observation $i$ to cluster $k$ in $M$
[3.] Assign each observation to the closest cluster
[4.] Compute the new modes for each cluster (get new array $M$)
[5.] Repeat until no observation changes the cluster
[6.] Output: $M = k$ x $m$ array of cluster modes; $C = n$ x 1 vector of cluster assignments
```

For the algorithm to run, two input choices must be made by the researcher ex-ante: the values of the initial modes and the optimal number of clusters. The values chosen as initial starting points are likely to influence the results. Thus, it is essential to have a reasonable array of input modes. We follow Khan and Ahmad (2013) and determine the initial modes based on the distribution in the data. Using a data-driven approach, we find the optimal number of clusters to be four. We describe both input decisions in detail in Appendix B1.

Before we run the algorithm, we prune the data to drop criteria with high occurrence of NA or with almost complete compliance since these contain little information for clustering.[4] In particular, we drop all criteria with more than 50% NA and all criteria with compliance rates over 95%, whereby NAs are ignored.[5] This leaves us with a remaining 30 criteria. Note that this procedure drops all critical criteria since the compliance rates for these are higher than 95%.

4.2 Results

Figure 1 and Table 1 show the result of the clustering exercise. The first panel of Fig. 1 presents the modes of each cluster where the criteria are listed on the horizontal and the

[4] Remember that occurrences of NA do not count toward any cluster assignment. Criteria with (almost) full compliance do not allow for any distinction between clusters.

[5] There are no criteria with compliance rates below 5%.

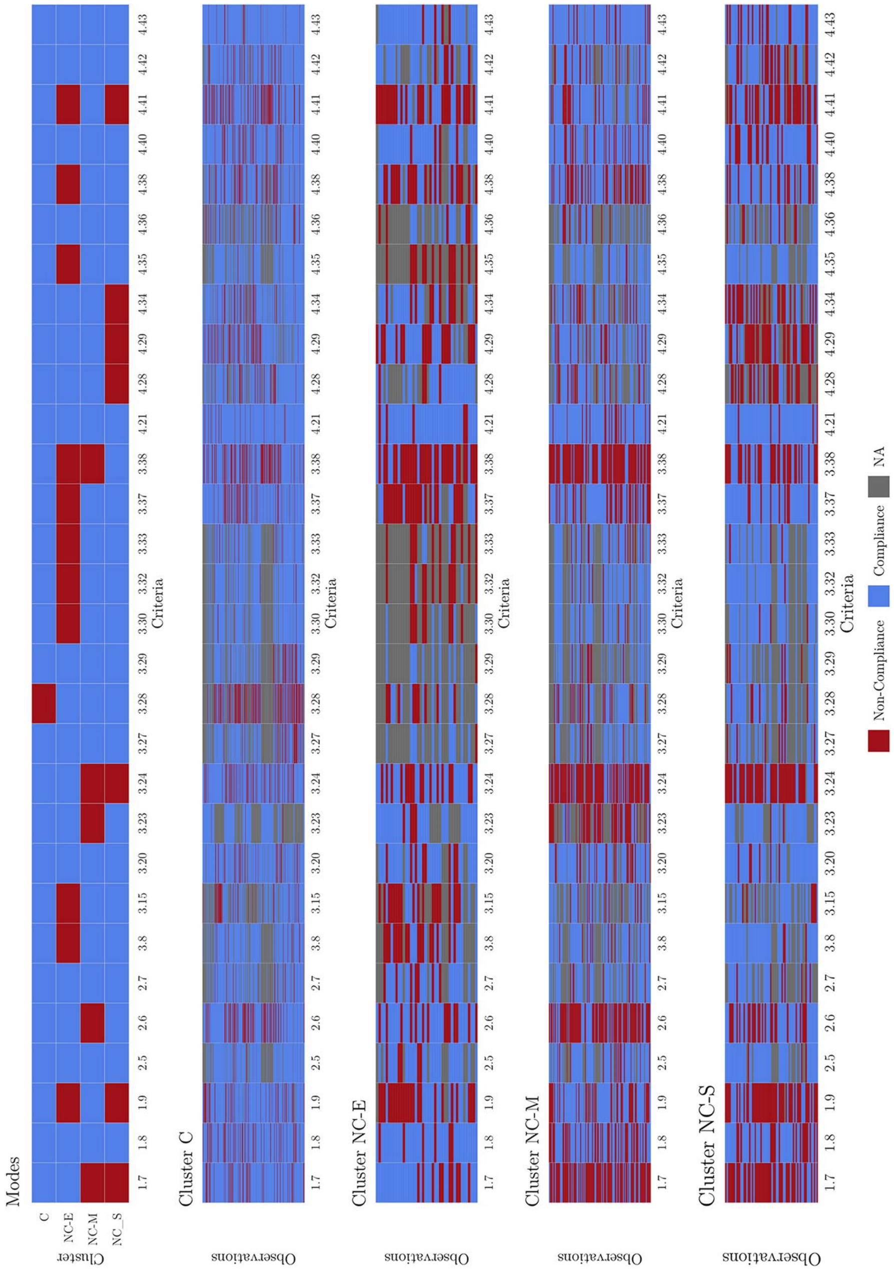

Modes
Cluster
C
NC-E
NC-M
NC_S
Cluster C
Cluster NC-E
Cluster NC-M
Cluster NC-S
Observations
Criteria
1.7
1.8
1.9
2.5
2.6
2.7
3.8
3.15
3.20
3.23
3.24
3.27
3.28
3.29
3.30
3.32
3.33
3.37
3.38
4.21
4.28
4.29
4.34
4.35
4.36
4.38
4.40
4.41
4.42
4.43
Non-Compliance
Compliance
NA

Table 1 Compliance shares by criteria and cluster

Criteria	All	C	NC-E	NC-M	NC-S
Principle 1: Effective Planning and Mgmt. System					
1.7	0.663	0.852	0.762	0.289	0.333
1.8	0.766	0.825	0.857	0.620	0.682
1.9	0.713	0.846	0.395	0.684	0.292
Principle 2: Biodiversity Conservation					
2.5	0.913	0.946	0.839	0.833	0.941
2.6	0.647	0.769	0.667	0.322	0.621
2.7	0.885	0.910	0.767	0.889	0.821
Principle 3: Natural Resource Conservation					
3.8	0.885	0.931	0.414	0.872	0.943
3.15	0.789	0.829	0.321	0.831	0.825
3.20	0.864	0.889	0.771	0.804	0.902
3.23	0.804	0.971	0.875	0.308	0.958
3.24	0.607	0.799	0.595	0.275	0.262
3.27	0.855	0.848	0.923	0.889	0.811
3.28	0.568	0.460	0.550	0.750	0.822
3.29	0.845	0.843	0.909	0.836	0.853
3.30	0.938	0.975	0.385	0.924	0.932
3.32	0.929	0.967	0.143	0.925	0.979
3.33	0.880	0.934	0.214	0.802	0.935
3.37	0.759	0.822	0.314	0.658	0.864
3.38	0.540	0.702	0.244	0.217	0.500
Principle 4: Improved Livelihoods and Wellbeing					
4.21	0.913	0.937	0.881	0.851	0.924
4.28	0.816	0.888	0.862	0.798	0.316
4.29	0.750	0.823	0.553	0.830	0.327
4.34	0.796	0.883	0.645	0.773	0.431
4.35	0.943	0.979	0.000	0.964	1.000
4.36	0.842	0.873	0.737	0.804	0.763
4.38	0.734	0.823	0.394	0.591	0.746
4.40	0.844	0.883	0.889	0.798	0.712
4.41	0.697	0.754	0.359	0.798	0.426
4.42	0.848	0.887	0.828	0.825	0.673
4.43	0.922	0.963	0.895	0.892	0.785
N	561	332	42	121	66

We exclude criteria with more than 50% of observations being NA or more than 95% of observations being compliant (ignoring NAs)

clusters on the vertical axis. The remaining panels show the same criteria on the horizontal and each observation in a cluster on the vertical axis. The figure gives a good impression on how neatly the observations are sorted into the clusters. When we look at Table 1, we see the compliance share for each of the 30 criteria by cluster. If all observations in a cluster complied with a certain criterion, the value in the table would be 1. A value of 0 represents perfect non-compliance. Of course, such extreme values are not to be expected since most certificates are unique in their compliance pattern. We interpret a value below 0.5 as non-compliance. (Mode is non-compliance.) The criteria are grouped into the four principles of the SAN-RA Standard. In addition, the first column shows the overall compliance shares by criteria. The values range from 0.540 to 0.943.[6] In the following, we interpret these results by describing each cluster in turn. A detailed description of non-compliant criteria for each cluster can be found in Appendix C.

4.2.1 Cluster C: compliers

The first cluster is the largest and makes up over one half of total producers (N = 332). These producers achieve very high compliance levels in almost all criteria. As can be seen in Fig. 1, the only exception is criterion 3.28 (vegetative barriers between pesticide-applied crops and areas of human activity), where a majority does not comply. This seems to be a very specific criterion, and from column 1 of Table 1, we learn that it has one of the lowest overall compliance levels. The fact that a large number of producers that are not compliant with this criterion end up in the cluster with the otherwise highest compliance suggests that it is barely correlated with other criteria.

4.2.2 Cluster NC-E: non-compliers with natural resource conservation (environment)

The second cluster describes a small category of producers (N = 42) that do not sufficiently implement environmental criteria. They disregard rules on the safe application of pesticides and face problems with water use and erosion as well as with their waste management. Additionally, workers do not benefit from sufficient health protection. Moreover, there is a high occurrence of management issues in these same areas, where careful planning and evaluation are lacking.

4.2.3 Cluster NC-M: non-compliers with management criteria

The third cluster comprises just under one quarter of producers (N = 121). A look at Fig. 1 reveals that these certificates have few criteria with non-compliances. However, a pattern appears: all criteria that are not complied with are related to the farm management or the group administrator. They do not develop a farm management plan, nor a plan to increase and restore the native vegetation, nor one for pest and waste management. Interestingly, these producers comply to a large extent with other criteria concerned with the actual task of conserving natural resources or implementing social criteria. The issues seem to lie solely with the (group) management.

4.2.4 Cluster NC-S: non-compliers with social criteria

Finally, the last cluster groups producers that have low compliance with social criteria. This group contains over one ninth of producers (N = 66). Workers suffer from insufficient

[6] Remember that criteria with compliance shares above 0.95 are excluded from the analysis.

housing and incomes. Also, the workers' representation is insufficient, i.e., there is no *Occupational Health and Safety committee*. Additionally, there are some problems with leadership such as the lack of upkeep and evaluation of the farm management plan as well as the recording of pest infections.

5 Characteristics of the clusters

So far, we have described the compliance patterns peculiar to each of the four clusters. There clearly seems to be a pattern that has some meaning in that each cluster has its own area of problems. In this section, we add more variables to the analysis to characterize the certificates in each cluster in more detail. This sheds light on who the producers in each cluster are and what could be potential drivers for non-compliance.

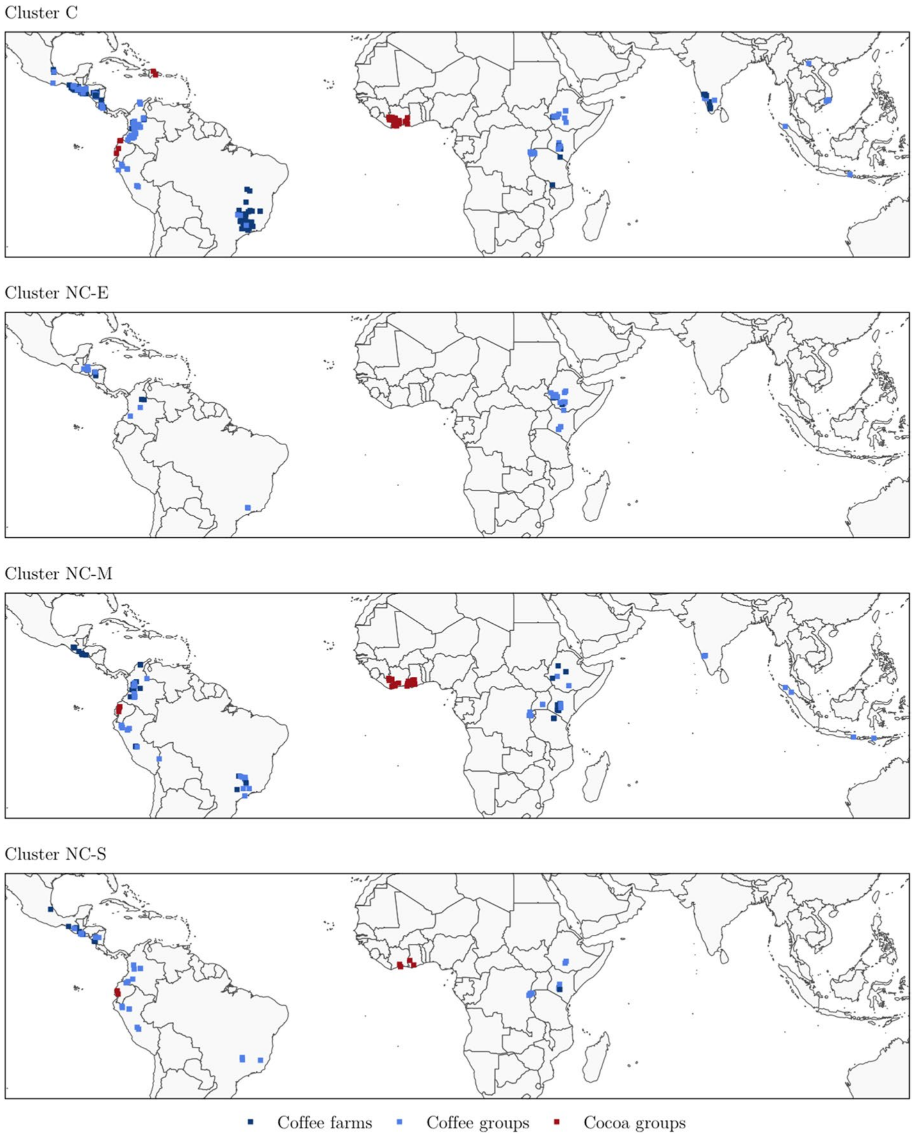
5.1 Type, crop, and location

Most empirical studies concerned with the effect of certification are looking at a geographically confined area and certain types of producers only. Thus, little is known about spatial heterogeneity of impact or about which crops are especially suited for certification. Figure 2 illustrates the location of the certificates per cluster and crop-type combination. Table 2 shows some additional indicators regarding location, type of producer, crop, and experience with certification. In cluster C, coffee production-especially on individual farms-is slightly over-represented. Figure 2 shows that most of these farms are located in Brazil. Unsurprisingly, producers in this cluster have on average been certified for longer and also had more previous audits. This either suggests a learning effect-at least in terms of how to pass the audit-or reflects a mechanical correlation if non-compliers drop out of the program over time. Producers in cluster NC-E are often located in East Africa, especially in Ethiopia, and consist nearly exclusively of coffee groups. Additionally, they have had the least amount of experience with certification. NC-M certificates are often found in cocoa group certificates. Nearly, all Ghanaian cocoa producers are assigned to this cluster. Also, producers in Central America rarely fall into this group. Finally, coffee group certificates located in Central America-especially in El Salvador-are over-represented in cluster NC-S.

5.2 Operation size

While the location and the type of producers provide some insights, we now turn to more specific indicators at the certificate level. Figure 3 provides statistics on the scale of production grouped by crop and certificate type. Cluster C consists of coffee farms that are large in area but employ relatively few workers. Nevertheless, they achieve the highest output per hectare, which suggests a high degree of mechanization. This high productivity corresponds to large-scale coffee farms, mainly in Brazil. Looking at coffee group certificates, the compliers tend to have fewer member farms, but the individual farms are larger, both, in terms of area and number of workers. It is plausible that both findings arise because large, productive producers face a smaller adaptive burden to comply with the standard.

For both, coffee farms and groups, producers in cluster NC-E struggle with productivity, especially in per worker terms. The group certificates consist of many small farms that


Cluster C

Cluster NC-E

Cluster NC-M

Cluster NC-S

Fig. 2 Maps of clusters. *Note*: To comply with the confidentiality agreement and to improve visibility, we dropped country-crop-type combinations with fewer than five observations from the figure

employ little labor from outside the household. This corresponds well with our findings in Sect. 5.3 below that the poorest farmers tend to be found in this cluster. Output per hectare is rather high for certificates in cluster NC-M, but compared to the complier cluster, similar amounts of output are generated by more workers for coffee farms, and by more member farms for coffee groups. Meaning the actual productivity, given inputs, is lower than for cluster C. The fact that the number of workers for farm certificates and the number of farms for group certificates are high might also make them more difficult to manage,

Table 2 Type and continent

	All	C	NC-E	NC-M	NC-S
Type					
Coffee	0.87	0.89**	0.98***	0.76***	0.91
Farm	0.41	0.48***	0.24***	0.36*	0.27***
Years certified	4.43	4.80***	3.22**	3.98*	4.21
Previous audits	0.15	0.19***	0.05***	0.11*	0.11
Continent					
Africa	0.27	0.18**	0.50***	0.44***	0.26
Asia	0.07	0.08	0.05	0.07	0.05
Central America	0.28	0.33**	0.26	0.08***	0.38**
South America	0.35	0.37*	0.19***	0.37	0.29

Cluster means along with the results of one-vs.-rest T-test. The symbols *, **, *** refer to statistical significance at the 10%, 5%, and 1%-level, respectively

which would explain the predominance of issues within these criteria. Contrary to the coffee groups in NC-E and NC-M, producers in cluster NC-S tend to employ some labor beyond the household members, which might explain why they struggle more with social criteria. The few cocoa groups in cluster NC-S achieve very high productivity with little labor input.

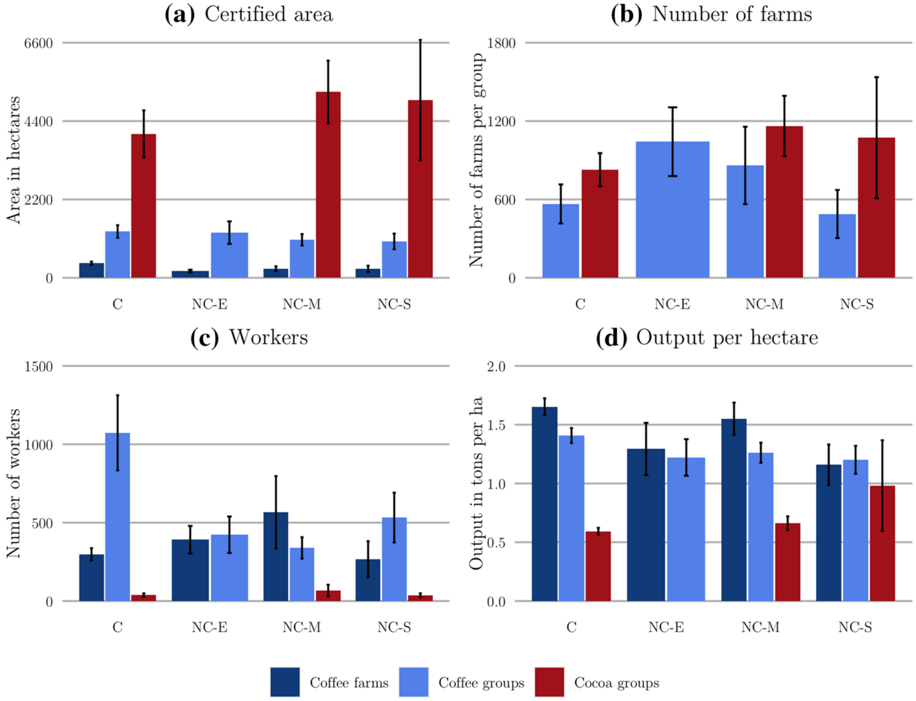


Fig. 3 Operation size. *Note* The black lines represent standard errors. We drop the single cocoa farm in the data and the single cocoa group in cluster NC-E from this analysis

Table 3 Economic, political, and geographic indicators

	All	C	NC-E	NC-M	NC-S
Economic and political development					
GDP p.c.	8792.59	9483.65***	5753.02***	8298.14	8145.82
Night lights	7.42	7.28	7.75	6.76	9.17*
Population density	487.13	416.92**	670.42	599.89	517.00
Lights per pop. dens.	0.03	0.03**	0.02***	0.02***	0.02*
Child mortality	26.35	24.24**	32.52***	31.30***	24.00*
Polity 2	6.28	6.62***	3.57***	6.35	6.14
Ethnic frac.	0.56	0.55**	0.59	0.64***	0.50***
Protected area	0.12	0.12	0.20**	0.10	0.12
Distance to city	71.84	66.55**	79.46	92.21***	56.28**
Market access					
City office	0.19	0.15***	0.24	0.20	0.33***
Ruggedness	166.77	172.81**	166.89	140.70***	184.13
Export cost	36.33	37.31**	32.34**	38.72*	29.58***
FDI	2.93	2.92	3.50***	2.91	2.67**
Access to credit	64.54	65.06	52.62***	63.64	71.14***
Agriculture					
Agricultural wage	327.93	349.76***	195.60***	301.91*	360.98
Agriculture % GDP	12.90	11.22**	20.54***	15.03***	12.56
Agriculture % exports	4.43	4.24**	5.58	5.31***	3.14***
N	561	332	42	121	66

T-test for one-vs.-rest. The symbols *, **, *** refer to statistical significance at the 10%, 5%, and 1%-level, respectively

5.3 Economic, political, and geographic factors

The economic and political environment plays a crucial role for the viability of VSS. On the one hand, VSS should target the vulnerable producers in poor regions to have the biggest impact (e.g., Dragusanu et al., 2014). On the other hand, certification is more likely to fail in difficult circumstances. Fundamental conditions for investments in certification are market access (e.g., Dammert and Mohan, 2015), a sound political setting (e.g., Naylor, 2014; AbarcaOrozco, 2015), the availability of skilled workers and credit (e.g., Kirumba and Pinard, 2010), and to some degree competitiveness with non-certified producers (e.g., De Janvry et al., 2015). VSS organizations aim to improve these conditions locally, and however, better initial conditions reduce the adaptive burden once certification is put in place and, hence, facilitate compliance.[7]

Table 3 offers an array of variables along these dimensions. The first section reports indicators on the economic and political development. Clearly, the development level is highest around cluster C producers. They have the highest GDP per capita, which is also reflected in night/time lights if put in relation to population density. Additionally, child

[7] This is sometimes referred to as the Sustainability Standards Paradox (e.g., Potts et al., 2014).

mortality is comparably low, a concept closely related to poverty and health. Finally, these producers' countries also fare best in the Polity II index, a measure for the level of democracy. Further, these producers are situated in countries, where agriculture is of minor importance in the overall economy and agricultural wages are quite high. These findings suggest that RA certification works best in already well-off places.

Producers in cluster NC-E are situated in much poorer countries with high population density, which is again reflected in night-time lights. Child mortality is high and the level of democracy a magnitude lower than in any other cluster. Agriculture on average still makes up over 20% of total GDP, and exports of agricultural commodities remain important. It is noteworthy that it is the environmental and not the social criteria that are neglected in the poorest places. An explanation could be that poorer producers often do not have the means to implement criteria regarding environment. In contrast, they are often small producers with few employees beyond the household members where complying with social criteria is less of an issue.

Clusters NC-M and NC-S are situated in contexts of intermediate levels of development. With similar income levels, cluster NC-M displays higher child mortality. This is in line with the longer distance to cities and high export costs implying a remote, rural setting. The high ethnic fractionalization is probably driven by the large share of West African producers in this cluster. Cluster NC-S producers are the most "urban" in that they are closer to cities, more likely to have city offices and face low export costs despite hilly terrain.

Table 4 Agricultural surroundings

	All	C	NC-E	NC-M	NC-S
Coffee certificates					
NDVI	202.25	203.41	197.57	200.96	201.76
Coffee area	309.38	336.19**	154.72***	261.23*	357.49
Coffee production	252.74	278.92**	114.76***	211.84*	281.43
Coffee yield p.h.	0.76	0.83**	0.57***	0.67***	0.72
Certification coffee	141.79	168.64**	46.22***	65.60***	192.36
Cocoa certificates					
NDVI	210.82	209.57	–	213.77*	204.49
Cocoa area	318.66	327.09	–	333.44	193.83**
Cocoa production	223.68	253.81*	–	206.08	117.91*
Cocoa yield p.h.	0.61	0.66*	–	0.58	0.41*
Certification cocoa	2.96	3.29	–	2.85	1.45*
Livestock density					
Cattle density	3651.86	3395.35**	5712.67***	3641.04	3681.83
Goat density	1107.50	668.51**	1611.44	1839.44*	1660.80*
Sheep density	855.32	596.91**	1729.97**	1202.59*	975.16
Pig density	1774.61	2059.99**	726.36***	1329.11*	1807.02

T-test for one-vs.-rest. The symbols *, **, *** refer to statistical significance at the 10%, 5%, and 1%-level, respectively. There is only one cocoa certificate in cluster NC-E, which we drop from the analysis

5.4 Agriculture

To get an impression of the agricultural landscape surrounding certified producers, we turn to several grid cell level variables reported in Table 4. We group the certificates into coffee and cocoa producers since the optimal growing conditions differ for the two crops. For both crops, it immediately becomes apparent that the compliers (cluster C) are located in places where the relevant crop is cultivated and certified intensively with high yields per hectare. The most probable interpretation is that these are locations that are well-suited for the growth of the crop.

For cluster NC-E, we find the lowest intensity of production and yield per hectare among coffee producers. Additionally, certification of coffee does not seem to be common in the locations of these certificates. Thus, the story might indeed be that farmers in less suitable places fail to comply with criteria regarding the environment. Similarly, producers in cluster NC-M are also located in areas that are not optimal for coffee production, as the yield p.h. is significantly lower and certification is not widespread.

Producers in cluster NC-S show different results depending on the crop. While the locations of the certificates seem rather suitable for coffee, the opposite is the case for the few certificates in NC-S producing cocoa. Surprisingly, these are the same cocoa producers that showed high output per hectare compared to the other clusters in Sect. 5.2.

The last panel of Table 4 reports densities of the four most common domestic animals (per square kilometer). Livestock density-with the exception of pigs-is significantly higher for cluster NC-E. However, the reasoning behind this is not clear. It could be that livestock density, especially through manure production, can have negative impacts on the environment. Alternatively, high livestock density could be an indicator for poorer regions, where farmers do not specialize in one specific crop but diversify by farming both, livestock and crops.

6 Classification

As an additional exercise, we use the information gathered in the previous section to predict cluster affiliation of new producers. We use one of the most popular machine learning procedures: classification. Specifically, we use a type of random forest algorithm by Strobl et al. (2009). To avoid overfitting, we take our results from the previous section and include only variables as predictors that have shown to be highly significant for cluster differentiation (at the 1%-significance level). Due to the imbalance of our clusters, we follow Delgado and Tibau (2019) and use the Matthew's Correlation Coefficient (MCC) as our performance measure. The classification method is described in detail in Appendix B2.

Table 5 Confusion matrix

	True			
	C	NC-E	NC-M	NC-S
Predicted				
C	35.62	1.63	5.74	3.89
NC-E	3.98	3.71	2.33	0.86
NC-M	11.82	1.60	10.39	3.18
NC-S	7.76	0.55	3.12	3.84

Percentage scores shown. The 16 numbers add up to 100%

6.1 Results

Table 5 shows the confusion matrix with cell counts in percent. The overall share of correct predictions is 0.54, and the MCC is 0.28. While the algorithm predicts cluster affiliation rather well for cluster C, it struggles with identifying the other clusters. There is also a high level of heterogeneity in accuracy depending on crop-country-type combinations, as shown in Fig. 4. Unsurprisingly, country-crop-type combinations where most observations are in the same cluster are well predicted. This is the case for Brazilian coffee farms, Guatemalan coffee farms and groups, or Indian coffee farms. We find that prediction is particularly weak in countries where the cluster algorithm does not yield clean results. For example, Kenyan coffee producers tend not to fit well in any of the four clusters and are consequently classified with very low accuracy.

Figure 5 shows the ten most important variables for the classification. Variable importance is the mean decrease in accuracy of the forest when randomly re-shuffling the values of a variable (also referred to as permutation importance). As we can see, geographic variables are crucial to determine cluster affiliation. Additionally, the type of certificate and the number of years certified seem to play an important role.

7 Summary and discussion

In this section, we present a summary of the results found in our analysis of compliance and then put them in perspective regarding limitations and existing literature.

We were able to find sensible clusters of non-compliance that allow to group producers according to different areas they struggle to comply with. Hence, our first hypothesis is fulfilled. Besides producers that are very successful in implementing the standard, we identify three clusters that have compliance issues in the areas of natural resource conservation, management, and social issues. Additionally, we can characterize the clusters to find out what type of producers belongs to which cluster. Unsurprisingly, the compliers are more predominantly found in more developed regions of Central and South America and tend to be large individual farms that achieve high yields per hectare-a hint toward the high mechanization prevalent on these farms. For groups, it is those with relatively few but large member farms that show high productivity. Interestingly, for coffee producers, it is an asset to employ a lot of workers, while for cocoa cultivators, the opposite holds. Finally, producers in the compliers cluster have been certified for longer suggesting a learning effect.

Producers in the second cluster-non-compliance with environmental criteria-have the least experience with certification and tend to be located in poor regions that are also less suitable for cultivation. Most of the certificates are coffee producers that are certified in groups and employ few workers outside of their household. Producers in the third cluster-management issues-are often found in cocoa production in West Africa. Most Ghanaian producers are assigned to this cluster. They have the highest share of ethnic fractionalization. Finally, issues with social criteria are prevalent in group certificates for coffee production in Central America, mainly in El Salvador. In line with these results, we find in our classification exercise that the variables most important to predict cluster affiliation are mostly of geographic nature, next to the number of years of certification and the type of certificate

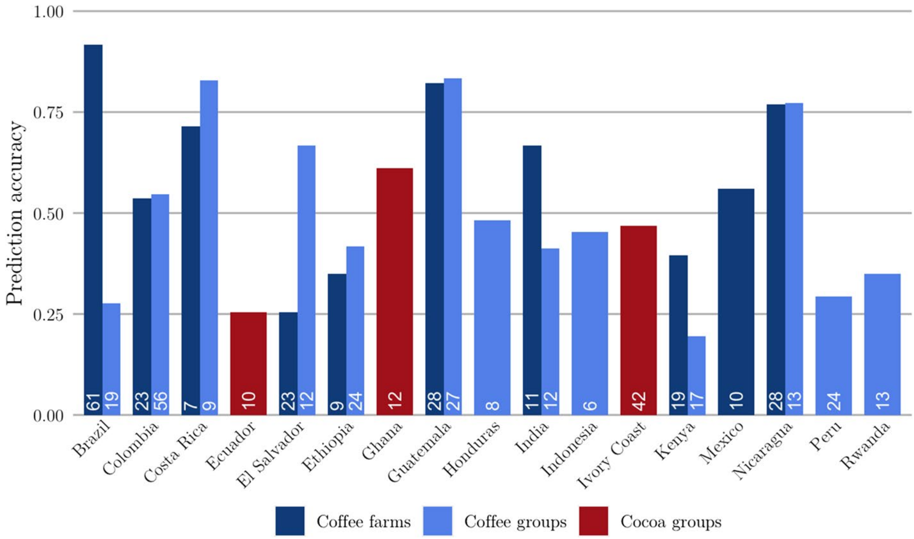


Fig. 4 Share of correct predictions *Note*: The white numbers at the bottom of each bar indicate the number of observations in that country-crop-type combination. We drop all country-crop-type combinations with $N \leq 5$

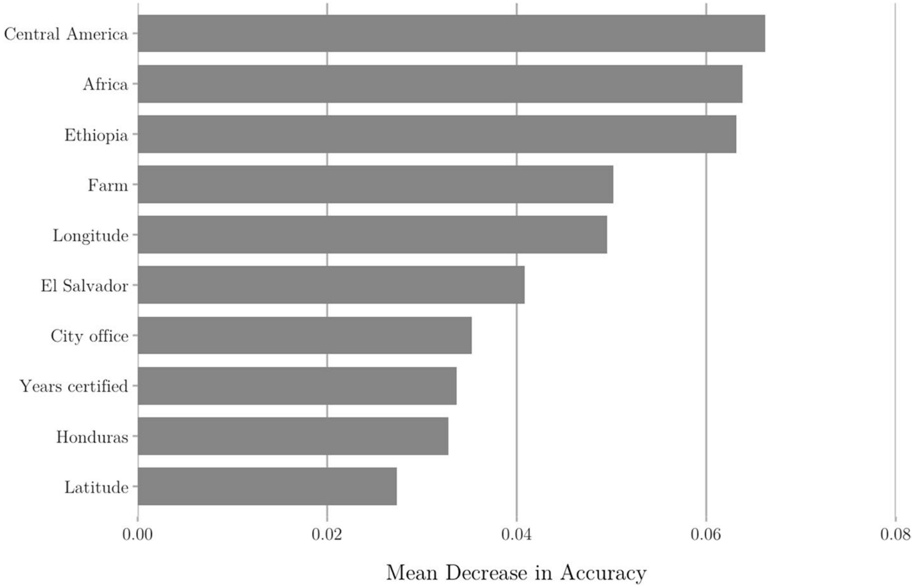


Fig. 5 Variable importance *Note*: The figure shows mean decrease in accuracy if a variable is removed from the model for the ten variables with the largest decrease

To be fair, the prediction accuracy of our classification exercise is modest at best. Thus, our second hypothesis regarding the identification of potential drivers of cluster affiliation is only partially fulfilled. Nevertheless, we are confident in recommending the approach for further research for several reasons. First, the number of observations is relatively small for classification, affecting the power of the analysis. Unfortunately, this is near-impossible to overcome given the present scale of RA certification. Second, we suspect that some critical information needed to accurately classify all observations is missing. To address this issue, more certificate-level data would need to be collected. And third, clustering all certificates world-wide bears the risk that the analysis is inept for certain crop-country combinations that do not fit in any of the clusters so obtained.

Comparing our results to the previous literature is difficult as the literature on compliance is extremely sparse. The only studies that we found focus on compliance of Brazilian coffee producers and compare group certificates to certified individual farms. Pinto et al. (2014) find similar overall compliance scores for individual farm and group certificates in 2011. Groups perform somewhat better in working conditions and health and safety measures, while individual farms outperform groups in wildlife and water protection as well as integrated crop management. Maguire-Rajpaul et al. (2020), who use compliance data from 2006 to 2014, find that Brazilian group certificates have somewhat higher non-compliance with a selection of management criteria and slightly lower social performance-although the differences are not statistically significant. Additionally, they find that the larger the certificates are in terms of area, the more compliant they are with social and management criteria. We can confirm the latter results for coffee even on a global level, as producers with a large certified area are over-represented in the compliers cluster. Further, our analysis suggests that individual farm certificates outperform group certificates in Brazil, and however, it is well possible that this difference only came about more recently, e.g., through improvements on large coffee farms. Talking to experts at RA, we find that certified Brazilian coffee farms are in many ways exceptional. They are much larger and more mechanized than the typical farms working with RA and contribute heavily to the global output of RA-certified coffee.[8]

8 Implications

In the introduction, we have outlined the importance of our research for two main avenues: highlighting the issue of heterogeneous compliance for causal inference and the improvement of VSS through systematic assignment of tailored training units and ex-ante risk assessment of new producers. Hence, in this chapter, we want to discuss implications of our results for both standard organizations and researchers. Previous research on the impact of certification with VSS was mostly concerned with selection bias. The rationale behind this is that producers with different characteristics differ in their take-up of certification. In Sect. 4, we have shown that certified producers additionally differ considerably in compliance, which suggests that even if treated, the outcome may vary among producers. Researchers interested in specific outcomes of certification are therefore encouraged to take into account compliance in addition to selection bias. For treated, i.e., certified, producers, the level of compliance with the VSS can potentially be observed by consulting audit data. However, this is never possible for control groups. One possible way to address this issue is to predict compliance and match to each treated observation a control unit with

[8] Brazilian farms make up about one sixth of producers in cluster C. We repeated the whole analysis (clustering, description, classification) without them and could not find major differences. Both the clustering and the main conclusions continue to hold.

similar expected compliance if it would certify (e.g., using a random forest algorithm). To do this, one needs to know the drivers of compliance with the VSS. In Sects. 5 and 6, we have identified several variables that possibly spur compliance. These insights are a starting point to guide future research in the adequate choice of the sample and matching technique.

Further, past impact studies were mostly concerned with very specific settings, e.g., small-holder coffee growers in Ethiopia. Compliance patterns for some potential settings such as coffee farms in Brazil or Guatemala are well predicted by our global study. In other settings like cocoa groups in Ecuador or Ivory Coast, it does considerably worse. However, it should be possible to repeat the analysis for a specific context to generate adequate clusters and significant predictors to be used in a particular causal study. In this sense, our analysis provides an example for a procedure to address this important issue.

Regarding implications for standards organizations, the potential for risk assessment remains limited in the face of the data currently available. While it is possible to predict compliance patterns of some larger homogeneous groups quite accurately, it is near impossible for others. We argue that this is due to missing information. Admittedly, prediction of clusters is further complicated by the ill fit of some of the observations to any of the clusters. However, we are able to identify those observations and we know where the prediction is likely to be accurate and where less so. Additionally, for the cases where classification is expected to be difficult, there is the option to turn to the full probability table of the prediction outcome. The classification algorithm returns a probability of belonging to each cluster for every observation. If we allow for some degree of human judgment, this information can be very useful in forecasting the compliance issues of a newly certified producer.

Finally, the clustering exercise in Sect. 4 lays a useful foundation for the systematic assignment of training units to certified producers in order to enhance their compliance with the standard. Our analysis would propose four different specifically tailored training courses for certified farmers in accordance with the four identified clusters. On the one hand, this would allow to bundle resources and to develop an education program to its highest effect. On the other hand, together with an improved prediction or insights from a first audit, it would become straightforward to assign each certified producer to the best training. While this approach seems efficient, it has one major shortcoming: It does not provide a solution for outliers within a cluster. However, there is a trade-off. A small number of training units can efficiently be handled at the cost of having misfits within the groups. In contrast, a large number of training units (equal to the number of certificates in the extreme) would allow for made-to-measure guidance with the expected increase in costs.

9 Conclusion

Voluntary sustainability standards are gaining popularity among conscious consumers. However, the effects of certification are still not well understood. We contribute to this scientific debate by investigating compliance-a complex and to date understudied aspect of the workings of VSS. For one particular voluntary sustainability standard-Rainforest Alliance-we show how compliance patterns can be clustered in order to understand and simplify the problem. We further identify possible drivers of these patterns, which can potentially be used for sophisticated causal inference studies, but also to improve VSS through targeted training or better risk assessment. Finally, we run a random forest algorithm using these drivers as predictors. Even though we find only a mediocre accuracy of the classification, the approach is promising since more data will be collected in the future. Thus, our study, while providing first descriptive insights into the complex matter of compliance, also outlines an avenue for further research toward understanding the mechanisms of certification.

Appendix A: Data

Certification by other standards

It might be important whether a producer operates in an area where certification is common. Thus, we include the number of certificates issued to producers in a certain area. These data were constructed by Tayleur et al. (2017) who map all certificates by 120 standards organizations by crop. We use the data for coffee and cocoa. The resolution is 30 x 30km.[9]

Child mortality

We use child mortality as a proxy for poverty and health. The Center for International Earth Science Information Network (2018a) provides a map with a 1 x 1km resolution. The source is sub-national administrative data. The values range from 0 to 176.6 deaths per 1000 live births.

Distance to city

To capture market access of producers, we use the travel distance to the next large city. Weiss et al. (2018) construct a dataset with the travel time to the next urban center with 50,000 or more inhabitants and a population density of at least 1500 people per square kilometer. The data are in minutes driving time by car, and the resolution is 1 × 1 km.

Harvest data

Monfreda et al. (2008) provide crop maps that report area, output, and yield per hectare for all major crops. The data are averages over the years 1997 to 2003 and are based on survey data. These data proxy the suitability of an area for a certain crop and how commonly this crop is grown in that area. The resolution is approximately 10 x 10km.

Livestock

We include livestock density for the four most important domestic animals: cattle, goats, sheep, and pigs. Data for 2010 are available from the Food and Agriculture Organisation (2010). The measures are in animals per square kilometer, and the resolution is 10 × 10 km.

[9] We report the resolution approximately in kilometers. However, the data are in decimal degrees, which are not the same length everywhere on the planet.

Night-time lights

Night-time lights have been shown to be a good proxy for economic activity (see, e.g., Henderson et al., 2011; Bruederle and Hodler, 2017). We obtain the satellite data for the year 2013 from the National Oceanic and Atmospheric Administration (2020). They have a resolution of roughly 1 × 1 km and indicate the light density at night. The values range from 0 to 63 where 63 indicates the brightest light. Note that the data are bottom as well as top coded due to the sensitivity of the satellite's sensor. We remove all large water surfaces from the data in order not to reduce the mean in the buffer for producers close to the coast.

Population

A key indicator is population density. The Center for International Earth Science Information Network (2018b) reports population density on a 1 × 1 km resolution. The numbers are people per square kilometer in 2015.

Protected areas

Since many criteria of the RA standard are related to environmental protection, we want a measure for the prevalence of protected areas around producers. UNEP-WCMC (2017) map all protected areas on a 1 × 1km resolution. The values are 1 for protected areas and 0 otherwise. The data are for 2015.

Ruggedness

To capture the topographic circumstances, we use the terrain ruggedness index (TRI), which was originally devised by Riley et al. (1999) and made available by Nunn and Puga (2012). If $e_{r,c}$ is the elevation in meters of row r and column c of a grid, then the TRI is calculated as $\sum_{i=r-1}^{i=r+1} \sum_{j=c-1}^{j=c+1} (e_{i,j} - e_{r,c})^2$. The resolution is 1 x 1km.

Vegetation

The normalized difference vegetation index (NDVI) measures the density of green vegetation on the earth surface (Huete et al., 1999). The resolution is 10 × 10 km, and the values range from 0 (no green) to 255 (very green). The highest value is assigned only to water surfaces, which appear very dark. We exclude these values from the dataset. The data are for 2019 and report yearly averages.

Country level data

To gain additional insights, we add a number of variables on country level.[10] From the World Bank's *World Development Indicators*, we obtain data on GDP and GDP per capita in 2011 I$, the share of agriculture in GDP and in exports, and foreign direct investment (FDI) as a share of GDP. From the *Doing Business Indicators*, we obtain a measure for access to credit and the cost of exports. The latter is computed as the mean

[10] All data were retrieved on April 1, 2020.

cost of documentation and border compliance. We use the score, which we recode such that it ranges from 0 (very cheap) to 100 (very expensive). The same range applies for access to credit (0 = very hard, 100 = very easy). All these data are for 2018. From the International Labor Organization (ILO), we retrieve data on agricultural wages in 2011 I$. Since not all countries have an estimate for all years, we chose to use the latest available estimate, not earlier than 2010. Finally, from the *Quality of Government* dataset, we use the indicator for ethnic fractionalization and the Polity II index, which measures the level of democracy from −10 (autocracy) to +10 (democracy) (Teorell et al., 2020).

Appendix B: method

B1: Clustering

Initial versus resulting modes

The approach of Khan and Ahmad (2013) is built on the insight that some observations are very similar and, therefore, are likely to share the same cluster irrespective of the choice of initial modes. More specifically, we find the frequency of each compliance pattern in the data. We then choose the patterns that occur more than once and use hierarchical clustering to reduce the number of clusters to the prespecified number k. In hierarchical clustering, the two closest clusters (or observations) are combined according to the distance measure (1), where the maximum distance between any two individual components of two clusters is used. This step is repeated until we have only k clusters. Thus, our initial modes represent the compliance patterns (and their similar neighbors) that are most frequent in the data. Figure 6 shows the initial and resulting modes for the clusters.

The number of clusters

We use two common goodness of fit measures to determine the optimal number of clusters: the *within cluster sum of distances* (WCSD) and the *silhouette coefficient* (SC). The WCSD is computed as the sum of the distances (1) between each observation in a cluster and the corresponding cluster mode. Thus, we want the WCSD to be as small as possible to make sure that observations within a cluster are similar to each other. However, the WCSD decreases in k by construction. Hence, we use an *Elbow Test* where we graphically look at this decrease and choose the k where there is a kink. At this point, there is no more meaningful reduction in the WCSD but solely the mechanical one. Figure 7 shows WCSD for $k \in [2, 10]$. As can be seen, it is everything but well-behaved. Candidates for k are 4 but also 7.

The SC measures how well every observation fits into their assigned cluster by looking at both cohesion and separation (Rousseeuw, 1987). Hence, it does not only take into consideration how small the distances are within the assigned cluster (cohesion), but also how far away the observations are from the observations in the nearest neighboring cluster (separation). Concretely, the silhouette value for each observation is calculated as

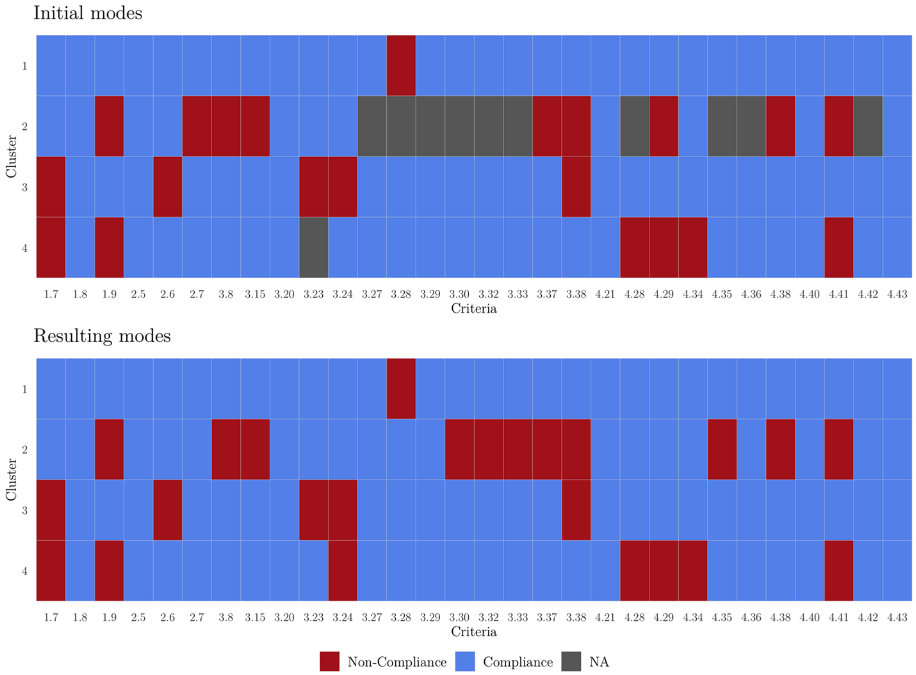


Fig. 6 Initial and resulting modes of clustering algorithm

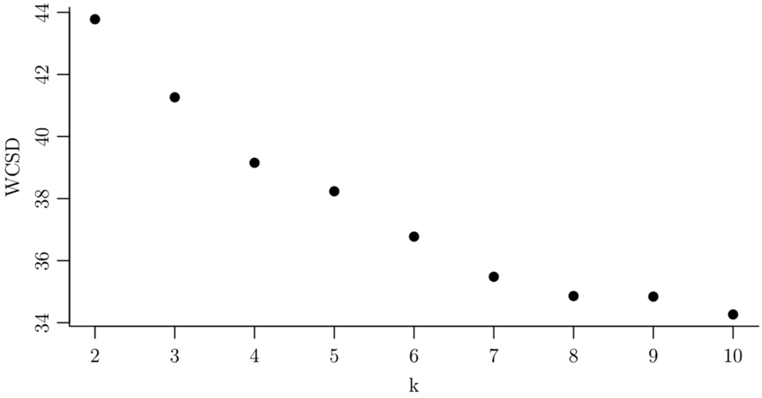


Fig. 7 Within cluster sum of distances. *Note*: Within cluster sum of distances (WCSD) for different numbers of clusters (k). We observe slight kinks at k = 4 and at k = 7

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{3}$$

where $b(i)$ is the smallest average *between* cluster dissimilarity, and $a(i)$ is the average *within* cluster dissimilarity. The silhouette coefficient is the mean of all silhouette values and ranges from -1 to 1. Clearly, we want the SC to be as high as possible. Figure 8 shows the silhouette coefficient for $k \in [2, 10]$. We can see that there is a big drop after $k = 5$.

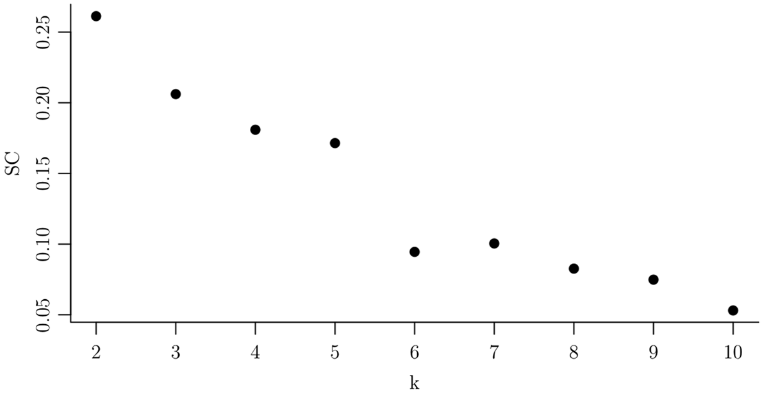


Fig. 8 Silhouette coefficient. *Note*: Silhouette coefficient (SI) for different numbers of clusters (k). We observe a large drop after k = 5

Thus, we will not consider more than five clusters for our analysis. Combining our insights of the two measures, we choose four as our optimal number of clusters.

B2: Classification

In a first step, we compare different algorithms by evaluating their performance. The most commonly used performance measure is accuracy, the ratio of correctly predicted observations divided by all predicted observations. We decide against using accuracy as our main performance measure due to the imbalance of our classes (although we still report it in Table 6). The biggest class, cluster C, is roughly eight times the size of our smallest class, cluster NC-E. If we do not control for this issue, the algorithm will predict cluster C too often to maximize accuracy. Therefore, we follow Delgado and Tibau (2019) and judge the algorithms based on a different performance measure: Matthew's Correlation Coefficient (MCC). MCC is essentially a coefficient for the correlation between the true and the predicted classes. Additionally, we resample the data for the training. We use upsampling, where observations of the minority classes are randomly duplicated until all classes have the same size as the majority class. The advantage compared to down-sampling is that there is no information loss which is ideal for classification with few observations.

Table 6 Selected performance measures for different algorithms

Measure	ctree	Glmnet	rf	cforest
MCC	0.186	0.244	0.267	0.279
	0.061	*0.065*	*0.060*	*0.055*
Accuracy	0.412	0.493	0.557	0.536
	0.058	*0.043*	*0.035*	*0.037*

Standard errors reported in italics. *ctree* refers to the conditional inference tree in the *party* package. *glmnet* is based on logistic regression with elastic net penalty. *rf* is the standard random forest from *randomForest* package. *cforest* is described further below

In Table 6, we present two performance measures, MCC and accuracy, for some of the most commonly used algorithms. To avoid overfitting by adding all available data, we utilize our results from the previous part to decide which variables to include in the classification. We use all variables that have been highly significant ($p < 0.01$) in the one-vs-rest t-test for at least one of the clusters. Additionally, we add the full location data and exclude country-level data since this information is captured by the country dummies. For all algorithms, we adopt repeated fourfold cross-validation with one hundred repetitions for more robust results. Looking at our preferred measure, MCC, it becomes clear that the *cforest* algorithm performs best and we will therefore execute the classification with this algorithm.

The *cforest* algorithm ships with the *party* package in *R* (for a brief overview, see Strobl et al., 2009). It is based on the random forest algorithm introduced by Breiman (2001). However, it uses conditional inference trees as developed by Hothorn et al. (2006) as base learners. In line with the random forest algorithm, it generates 500 (default) trees using only a limited number of variables to generate the splits at each node. This number is tuned to obtain the best fit (16 in our case). Further, the trees are not grown to their maximal size as the leaves need to contain at least seven (default) observations. In contrast to the standard random forest-where each tree is grown using a random sample with replacement (called bootstrap aggregation or bagging)-we follow Strobl et al. (2007) who suggest using sampling without replacement in combination with conditional inference trees. According to their approach, a random subset of the original sample containing 63.2% of observations is randomly chosen for each tree in the random forest.[11] The particularity of conditional inference trees as opposed to conventional trees is that splits are based on a significance test using a permutation procedure. This approach makes sure that those variables that have the strongest association with the outcome variable are selected to grow a tree. A great advantage is that it corrects for a bias in the variable selection toward those that have many possible splits (e.g., variables with many categories or continuous variables) (Hothorn et al., 2006). Additionally, in *cforest*, a probabilistic aggregation is used with leaf size as weights instead of majority voting which is used in the standard random forest procedure (Hothorn et al., 2004).

Appendix C: Non-compliant criteria by cluster

See Tables 7, 8, 9, 10.

Table 7 Non-compliance of cluster C

Principle 3: Natural Resource Conservation	
3.28	**Farms** establish and maintain non-crop *vegetative barriers* compliant with Rainforest Alliance parameters for vegetative barriers or Rainforest Alliance non-application zones *between pesticides applied crops and areas of human activity*.

[11] 0.632 corresponds to the expected fraction of unique observations that would end up in the standard random sample with replacement.

Table 8 Non-compliance of cluster NC-E

Principle 1: Effective Planning & Mgmt. System	
1.9	The **farm management** and **group administrator** analyze at least annually records on farm inputs and production to *evaluate the achievement of the farm management plan* and adjust the plan for the following year
Principle 3: Natural Resource Conservation	
3.8	**Farms** reduce *water and wind erosion* through practices such as ground covers, mulches, re-vegetation of steep areas, terracing, filter strips, or minimization of herbicide use
3.15	**Farms** comply with applicable law for the *withdrawal of surface or groundwater* for agricultural, domestic or processing purposes
3.30	All *pesticides* are stored in a safely locked storage facility. Only people trained in pesticide risks and management have access to the pesticide storage facility
3.32	Potentially affected persons or communities are identified, alerted, and warned in advance about *pesticide* applications and prevented from access to pesticide application areas
3.33	Empty *pesticide* containers and application equipment are triple washed, and the rinse water is returned back to the application mix for re-application. Empty pesticide containers are kept in a locked storage area until safely returned to the supplier or, if the supplier does not accept empty containers, they are cut or perforated to prevent other uses. Containers may be re-used only for the original contents and only when labeled accordingly
3.37	*Waste* storage, treatment and disposal practices do not pose health or safety risks to farmers, workers, other people, or natural ecosystems
3.38	The **farm management** and **group administrator** develop and implement a *waste management plan*
Principle 4: Improved Livelihoods and Wellbeing	
4.35	**Farms** implement Restricted Entry Intervals (REI) for persons entering *pesticide* application areas without PPE that are at least 12 hours or as stipulated in the product's MSDS, label or security tag. For WHO class II products, the REI is at least 48 hours or as stipulated in the product's MSDS, label or security tag. When two or more products with different REIs are used at the same time, the longest interval applies
4.38	Workshops, storage areas, and processing *facilities are designed for safe and secure storage of materials* and equipped and identified in accordance with the type of stored substances and materials, are clean and organized, and have sufficient light and ventilation, equipment for firefighting, and means to adequately remediate any substance or spillage of materials

Table 8 (continued)

4.41	The **farm management** and **group administrator** provide workers with *medical examinations* as specified in the Occupational Health and Safety plan. Workers have access to the results of their medical examinations

Table 9 Non-compliance of cluster NC-M

Principle 1: Effective Planning & Mgmt. System	
1.7	The **farm management** and **group administrator** develop and update regularly a *farm management plan* to optimize productivity, input use efficiency, and comply with this standard.
Principle 2: Biodiversity Conservation	
2.6	The **farm management** and **group administrator** *develop a map* that includes natural ecosystems and agroforestry canopy cover or border plantings with estimated vegetation coverage and estimated percentage of native species composition. If the farm or group of member farms have less than 10% total native vegetation cover or less than 15% total native vegetation cover for farms growing shade-tolerant crops, the farm management and group administrator develop and implement a *plan to progressively increase or restore native vegetation*.
Principle 3: Natural Resource Conservation	
3.23	In the case of groups, the **group administrator** develops an integrated *pest management (IPM) plan* for the group. The **group administrator** *trains and supports its members* to implement this plan on the member farms.
3.24	The **farm management** and **group administrator** *record pest infestations*.
3.38	The **farm management** and **group administrator** develop and implement a *waste management plan*.

Table 10 Non-compliance of cluster NC-S

Principle 1: Effective Planning & Mgmt. System	
1.7	The **farm management** and **group administrator** develop and update regularly a *farm management plan* to optimize productivity, input use efficiency, and comply with this standard.
1.9	The **farm management** and **group administrator** analyze at least annually records on farm inputs and production to *evaluate the achievement of the farm management plan* and adjust the plan for the following year.
Principle 3: Natural Resource Conservation	
3.24	The **farm management** and **group administrator** *record pest infestations*.
Principle 4: Improved Livelihoods and Wellbeing	
4.28	When the **farm management** and **group administrator** provide *housing* to workers, or workers with their families, this housing meets the following conditions: a) Beds are not arranged in more than two levels; b) Natural light during the daytime and artificial light for the nighttime; c) Natural ventilation that ensures movement of air in all conditions of weather and climate; d) Functional and effective fire wood smoke evacuation or ventilation mechanisms well maintained or repaired; e) Non-leaking windows, doors and roofs; f) At least one toilet for every 15 persons, one urinal for every 25 men, one washbasin for every six persons or per family; g) At least one shower per 10 persons, separated by gender; h) At least one large laundry sink for every 30 persons; i) Installed and maintained fire extinguishing mechanisms; j) Marked safety exits.
4.29	If a *living wage* benchmark is provided, the **farm management** and **group administrator** document and implement a living wage plan, to progress toward payment of living wage. In the absence of a living wage benchmark, the farm management and group administrator assess current access of workers and their families to health care and basic education and develop and implement a plan for providing access to these services.
4.34	An *Occupational Health and Safety (OHS) committee is chosen by workers* for farms or group administrators with 20 or more workers. The committee participates in or carries out regular OHS reviews and its findings and decisions are considered in the updating and implementation of the OHS plan. Committee decisions and associated activities are documented.
4.41	The **farm management** and **group administrator** provide workers with *medical examinations* as specified in the Occupational Health and Safety plan (see Critical Criterion 4.14). Workers have access to the results of their medical examinations.

Acknowledgements We thank Henk Gilhuis and Deanna Newsom at UTZ/Rainforest Alliance for providing the data.

Funding Open access funding provided by University of Luzern.

Data availability The audit data that support the findings of this study are available from UTZ/Rainforest Alliance, but restrictions apply to the availability of these data and they are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of UTZ/Rainforest Alliance. The source of all other (publicly available) data is clearly stated in Appendix A.

References

Abarca-Orozco, S. J. (2015). *Production and marketing innovations in Fair Trade and organic coffee cooperatives in the Córdoba-Huatusco corridor in Veracruz, Mexico (Dissertation)*. Iowa State University.

Borsky, S., & Spata, M. (2018). The impact of fair trade on smallholders' capacity to adapt to climate change. *Sustainable Development, 26*(4), 379–398.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bruederle, A., & Hodler, R. (2017). Nighttime Lights as a Proxy for Human Development at the Local Level. *CESifo Working Papers*, 6555.

Cao, F., Liang, J., Li, D., & Zhao, X. (2013). A weighting k-modes algorithm for subspace clustering of categorical data. *Neurocomputing, 108*, 23–30.

Center for International Earth Science Information Network (CIESIN). (2018a). *Documentation for the Global Subnational Infant Mortality Rates, Version 2*. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY.

Center for International Earth Science Information Network (CIESIN). (2018b). *Documentation for the Gridded Population of the World, Version 4 (GPWv4), Revision 11 Data Sets.* NASA Socioeconomic Data and Applications Center (SEDAC), Palisades NY.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review, 106*(5), 124–27.

Cramer, C., Johnston, D., Mueller, B., Oya, C., & Sender, J. (2017). Fairtrade and Labour Markets in Ethiopia and Uganda. *Journal of Development Studies, 53*(6), 841–856.

Dammert, A. C., & Mohan, S. (2015). A survey of the economics of fair trade. *Journal of Economic Surveys, 29*(5), 855–868.

De Janvry, A., McIntosh, C., & Sadoulet, E. (2015). Fair trade and free entry: Can a disequilibrium market serve as a development tool? *The Review of Economics and Statistics, 97*(3), 567–573.

Delgado, R., & Tibau, X. A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS ONE, 14*(9), e0222916.

Dietz, T., Estrella Chong, A., Grabs, J., & Kilian, B. (2020). How effective is multiple certification in improving the economic conditions of smallholder farmers? Evidence from an impact evaluation in Colombia's coffee belt. *Journal of Development Studies, 56*(6), 1141–1160.

Dragusanu, R., Giovannucci, D., & Nunn, N. (2014). The economics of fair trade. *Journal of Economic Perspectives, 28*(3), 217–236.

Dragusanu, R. & Nunn, N. (2018). The effects of fair trade certification : Evidence from coffee producers in Costa Rica. *NBER Working Paper*, 24260.

Food and Agriculture Organisation. (2010). Gridded livestock of the world database (GLW v3.1).

Glasbergen, P. (2018). Smallholders do not Eat Certificates. *Ecological Economics, 147*, 243–252.

Hainmueller, J., Hiscox, M. J., & Sequeira, S. (2015). Consumer demand for fair trade: Evidence from a multistore field experiment. *The Review of Economics and Statistics, 97*(2), 242–256.

Henderson, V., Storeygard, A., & Weil, D. N. (2011). A bright idea for measuring economic growth. *American Economic Review, 101*(3), 194–199.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674.

Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine, 23*(1), 77–91.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Research Issues on Data Mining and Knowledge Discovery* (pp. 1–8).

Huete, A. R., Justice C., & van Leeuwen, W. (1999). Modis Vegetation Index (Mod 13): Algorithm Theoretical Bases Document, Version 3.

Ingram, V., van Rijn, F., Waarts, Y., Dekkers, M., de Vos, B., Koster, T., & R., T., and A., G. (2017). *Towards sustainable cocoa in Côte d'Ivoire*. Wageningen Economic Research.

Khan, S. S., & Ahmad, A. (2013). Cluster center initialization algorithm for K-modes clustering. *Expert Systems with Applications, 40*(18), 7444–7456.

Kirumba EG & Pinard F (2010). Determinants of farmers' compliance with coffee eco-certification standards in Mt. Kenya region. *Joint 3rd African Association of Agricultural Economists (AAAE) and 48th Agricultural Economists Association of South Africa (AEASA) Conference, Cape Town, South Africa, September 19–23, 2010.*

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237–293.

Krumbiegel, K., Maertens, M., & Wollni, M. (2018). The role of fairtrade certification for wages and job satisfaction of plantation workers. *World Development, 102,* 195–212.

Maguire-Rajpaul, V. A., Rajpaul, V. M., McDermott, C. L., & Guedes Pinto, L. F. (2020). Coffee certification in Brazil: Compliance with social standards and its implications for social equity. *Environment, Development and Sustainability, 22*(3), 2015–2044.

Monfreda, C., Ramankutty, N., & Foley, J. A. (2008). Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles, 22*(1), 1–19.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106.

National Oceanic and Atmospheric Administration (NOAA) (2020). Version 4 DMSP-OLS Nighttime Lights Time Series. Boulder, CO: National Geophysical Data Center. https://www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html. Accessed January 2020.

Naylor, L. B. (2014). *Decolonial Autonomies: Fair Trade, Subsistence and the Everyday Practice of Food Sovereignty in the Highlands of Chiapas (Dissertation)*. University of Oregon.

Nunn, N., & Puga, D. (2012). Ruggedness: The Blessing of Bad Geography in Africa. *The Review of Economics and Statistics, 94*(1), 20–36.

Oya, C., Schaefer, F., & Skalidou, D. (2018). The effectiveness of agricultural certification in developing countries: A systematic review. *World Development, 112,* 282–312.

Pinto, L. F. G., Gardner, T., McDermott, C. L., & Ayub, K. O. L. (2014). Group certification supports an increase in the diversity of sustainable agriculture network-rainforest alliance certified coffee producers in Brazil. *Ecological Economics, 107*, 59–64.

Potts, J., Lynch, M., Wilkings, A., Huppé, G., Cunningham, M., & Voora, V. (2014). *The State of Sustainability Initiatives Review 2014*. International Institute for Sustainable Development (IISD) and the International Institute for Environment and Development (IIED).

Riley, S. J., DeGloria, S. D., & Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Science, 5*(1–4), 23–27.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Technical report.

Ruben, R., & Fort, R. (2012). The impact of fair trade certification for coffee farmers in Peru. *World Development, 40*(3), 570–582.

Sellare, J., Meemken, E. M., & Qaim, M. (2020). Fairtrade, agrochemical input use, and effects on human health and the environment. *Ecological Economics, 176,* 106718.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(25), 1–21.

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on! A new, conditional variable importance measure available in the party package. *The R Journal, 2,* 14–17.

Tayleur, C., Balmford, A., Buchanan, G. M., Butchart, S. H., Corlet Walker, C., Ducharme, H., Green, R. E., Milder, J. C., Sanderson, F. J., Thomas, D. H., Tracewski, L., Vickery, J., and Phalan, B. (2018). Where are commodity crops certified, and what does it mean for conservation and poverty alleviation? *Biological Conservation, 217*, 36–46.

Teorell, J., Dahlberg, S., Holmberg, S., Rothstein, B., Pachon Alvarado, N., & Axelsson, S. (2020). *The Quality of Government Standard Dataset, version Jan20*. The Quality of Government Institute: University of Gothenburg.

UNEP-WCMC (2017). *World Database on Protected Areas User Manual, 1.5*. UNEP-WCMC, UK.

Van Rijsbergen, B., Elbers, W., Ruben, R., & Njuguna, S. N. (2016). The Ambivalent Impact of Coffee Certification on Farmers' Welfare: A Matched Panel Approach for Cooperatives in Central Kenya. *World Development, 77*, 277–292.

Waarts, Y., Ingram, V., Linderhof, V., Puister-jansen, L., Rijn, F. V., & Aryeetey, R. (2014). *Impact of UTZ certification on cocoa producers in Ghana, 2011 to 2014*. LEI Wageningen UR.

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature, 553*(7688), 333–336.